A Computational Perspective on some Cognitive Illusions

Kenneth D. Forbus

Northwestern University

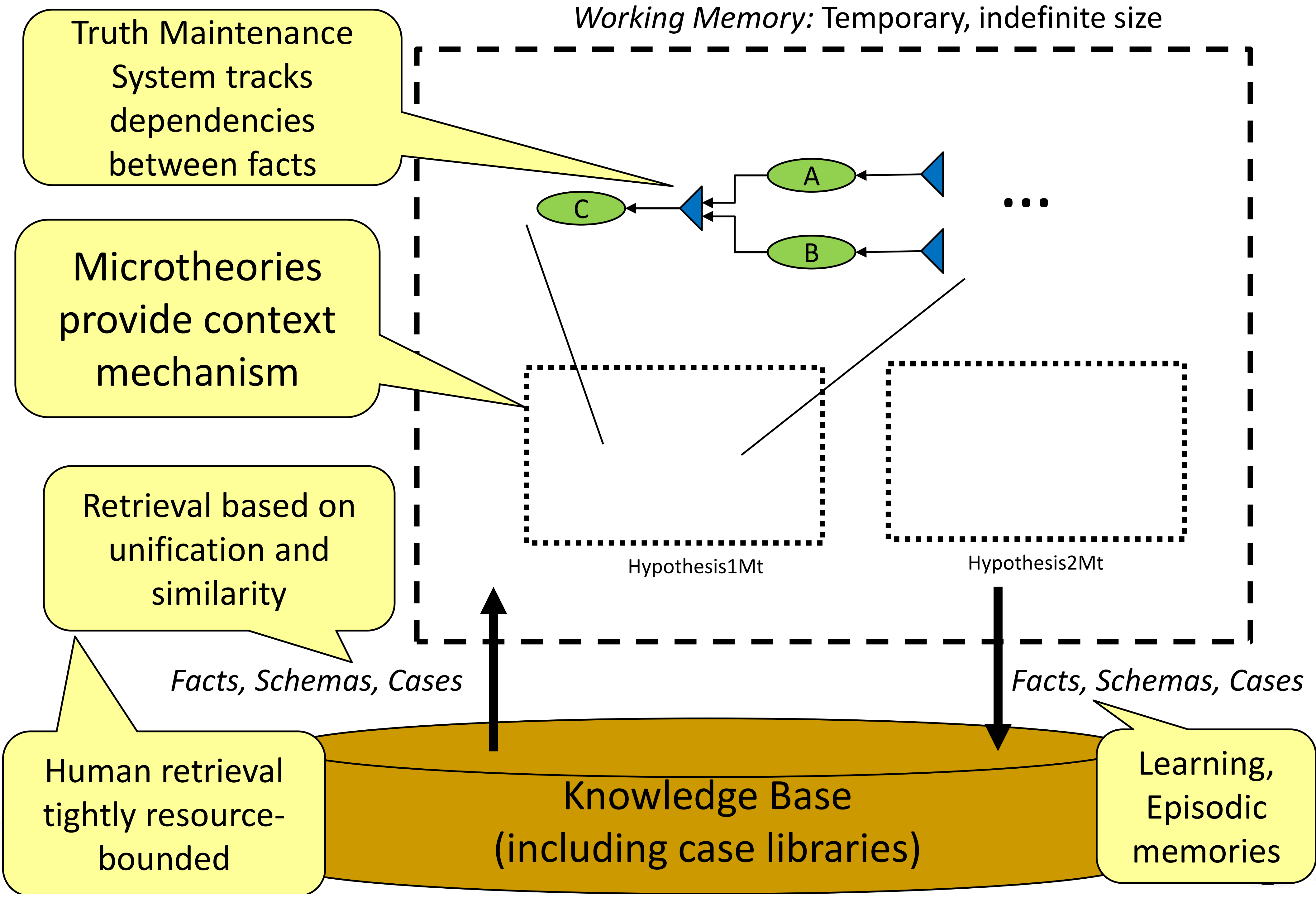
Cognitive Illusions

- Example: Confirmation bias
  - Many more (> 200 in Wikipedia)
  - Likely cost for flexibility of heuristic reasoning
  - They even happen with trained professionals
    - See Heuer’s *Psychology of Intelligence Analysis*
- Exact psychological mechanisms still under investigation
- Goal: Understand them computationally
  - Potential for more concise theory
  - Create AI reasoners that complement humans
  - More flexible AI reasoners
  - Understand what new cognitive illusions AI reasoners may have

Overview

- What they are ✓
- Notional reasoning architecture
- Three illusions
  - Confirmation Bias
  - Mirroring
  - Misinformation effects
    - With psychological prediction
- Related and Future Work

Simplified Reasoning Architecture



Working Memory: Temporary, indefinite size

Truth Maintenance System tracks dependencies between facts

Microtheories provide context mechanism

Retrieval based on unification and similarity

Facts, Schemas, Cases

Facts, Schemas, Cases

Human retrieval tightly resource-bounded

Knowledge Base (including case libraries)

Learning, Episodic memories

Illusion: Confirmation Bias

- Once a hypothesis is formed, people tend to
  - Pay attention to evidence that supports it
  - Ignore evidence that contradicts it
- Hypothesis: Confirmation bias is the cost paid for powerful human pattern-recognition abilities
  - e.g. similarity-based retrieval (MAC/FAC)
- Conjecture: AI systems can be arranged with similarly powerful pattern recognition abilities while avoiding confirmation bias
  - Example: Using Cyc-style microtheories to separate contexts eliminates interference in our similarity-based retrieval model

Illusion: Mirroring

- Believing that other actors think the way we do
- Hypothesis: Mirroring is the cost paid for the use of "like me" reasoning that may bootstrap human social reasoning and theory of mind
  - e.g. Rabkina's Analogical Theory of Mind
- Conjecture: AI systems can be arranged so that when cross-culture transfer is used, it is noted for further scrutiny
  - e.g. use of cultural stories to express protected values for moral decision-making (Dehghani et al. 2008)

Illusion: Misinformation Effects

- Example: Incidental incorrect facts in stories can override long-term knowledge
  - e.g. Rapp & Salovich, 2018
  - “[…] My boat was named after the mythical civilization that sank into the ocean, Pompeii. […]”
  - When asked what civilization sank into ocean,
    - Some answered “Pompeii”
    - Some answered “Atlantis”, but more slowly than those who did not receive misinformation
    - Some answered “Atlantis” with no speed difference

One Possible Explanation

- Facts stored with microtheories
  - Some evidence for this (Gerrig & Prentice, 1991)
- Individual differences in vetting
  - Heavy vetting on input: Pompeii either not stored or marked as incorrect (correct, no slowdown)
  - Vetting on retrieval: Question retrieves conflicting answers from WM & LTM (correct, slowdown)
  - No vetting: Incorrect answer retrieved from WM more quickly (incorrect)
- Similar effects can happen days later
  - Possible cause: Implicit analogical inference (Day & Gentner, 2007)
  - Novel prediction: Should get intrusion for similar, not just identical, materials

Related Work

- Lebiere et al 2013; Thomson et al. 2014 used ACT-R to model confirmation bias and several other biases
  - Matched against human data on sensemaking
  - Only worked with attribute-based models, not the kind of relational conceptual structures used here
- Models of mirroring
  - Kennedy et al. 2009; Hiatt et al 2011 used ACT-R for HRI experiments, embedded only, no cultural reasoning
- Unaware of prior models of misinformation effects

Future Work

- Build dataset(s) to support computational experiments
- Extend the set of cognitive illusions modeled
  - Work with psychologists to test new predictions
- Experiment with variations on the reasoning architecture to explore tradeoffs
  - Flexibility of human reasoning without the limitations?